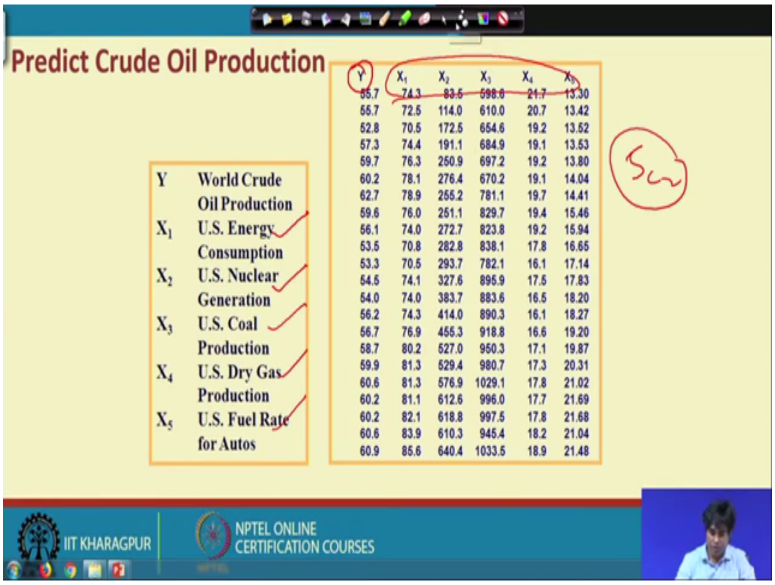
Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 24
Modelling Diagnostics (Contd.)

Hello everybody, this is Rudra Pradhan here, welcome to Engineering Econometrics. Today we will continue with Modelling Diagnostics, in that too discussion is on multiple entry problems. In fact, in the last lecture we have discussed this particular component and the issue is that it is the problem of linear relationship among the regressors, while regressing dependent variable with independent variables. So, one of the typical requirement is that a independent variable should be linearly independent but in reality with respect to you know real life problems, these independent variables are not actually completely independent.

So, what is our target here to check how you know how much the existence of relationship between or among these you know independent variables, so that means, technically we like to track out the degree of relationship among these independent variables. Whether it is a you know low and high you know it is not a issue but it should not be statistically significant. The moment the particular correlation coefficient will be statistically significant, then the estimated model which is derived through various mechanisms may not be good to you know use to predict or the kind of you know or to forecast any kind of you know engineering problem.

(Refer Slide Time: 02:01)



Predict Crude Oil Production

Y	World Crude Oil Production
X_1	U.S. Energy Consumption
X_2	U.S. Nuclear Generation
X_3	U.S. Coal Production
X_4	U.S. Dry Gas Production
X_5	U.S. Fuel Rate for Autos

Y	X_1	X_2	X_3	X_4	X_5
55.7	74.3	83.5	598.6	21.7	13.30
55.7	72.5	114.0	610.0	20.7	13.42
52.8	70.5	172.5	654.6	19.2	13.52
57.3	74.4	191.1	684.9	19.1	13.53
59.7	76.3	250.9	697.2	19.2	13.80
60.2	78.1	276.4	670.2	19.1	14.04
62.7	78.9	255.2	781.1	19.7	14.41
59.6	76.0	251.1	829.7	19.4	15.46
56.1	74.0	272.7	823.8	19.2	15.94
53.5	70.8	282.8	838.1	17.8	16.65
53.3	70.5	293.7	782.1	16.1	17.14
54.5	74.1	327.6	895.9	17.5	17.83
54.0	74.0	383.7	883.6	16.5	18.20
56.2	74.3	414.0	890.3	16.1	18.27
56.7	76.9	455.3	918.8	16.6	19.20
58.7	80.2	527.0	950.3	17.1	19.87
59.9	81.3	529.4	980.7	17.3	20.31
60.6	81.3	576.9	1029.1	17.8	21.02
60.2	81.1	612.6	996.0	17.7	21.69
60.2	82.1	618.8	997.5	17.8	21.68
60.6	83.9	610.3	945.4	18.2	21.04
60.9	85.6	640.4	1033.5	18.9	21.48

So, in order to you know knowing this particular you know issue, so let us start with the kind of you know component. So, this is a multiple regression problem, so, where we have a dependent variable and then we have independent variables. So that means, technically Y is the dependent variables and then we have 5 independent variables. So that means, the issue of multicollinearity will be you know starting you know very you know at a high rate you can say or the complexity will start increasing the moment you will one after another independent variables into the systems.

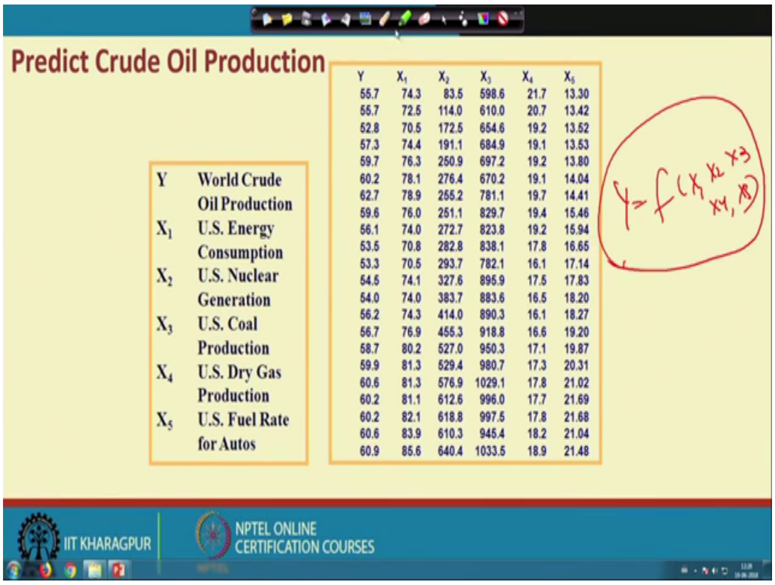
So, ultimately you know this is kind of you know problems, where we have one dependent variable with the 5 independent variables so; that means, we will have a different levels of you know relationship among the independent variable. Technically is you know there are couple of you know pairs will be find starting with X 1 X 2 X 1 X 3 X 1 X 4 X 1 X 5 and similarly X 2 X 3 X 2 X 4 X 2 X 5 then X 3 X 4 and X 3 X 5 and finally, X 4 X 5. So, ultimately, so, this is how the kind of you know case. So, its 5 C 2 you know times you will find such relationships.

And then out of all these relationships you know you will find some cases it is actually go against the you know model estimation or some case it may in favor you know model estimation. For instance you cannot say favor completely, but you know the technically the favorite of you know relationship or favorite of the model will come into the pictures, when the relationship among the regressions will exactly equal to 0, but in a real life scenario you will not find such kind of you know environment. So, anyway, so, this is a problem. So, where Y is the Y is the dependent variables and represents oil productions and it is affected by couple of you know independent variables like you know US energy consumption; that means, technically this oil production is at the word labels, total world oil productions and it is it is affected by couple of independent variables. If first one is US energy consumption, US nucleus generation, US coal production, US dry a dry gas productions and US fuel rate for automobiles.

So that means, technically means from this problem it is very much clear that you know the price of the production of you know oil production exclusively depends upon the us market, and that too with respect to energy consumption, nuclear generations, coal productions, gas productions and fuel rate. So, all these things are there to you know predict the oil price or oil productions, but the thing is that you know we are restricted to 5 independent variables and there is a high chance since you know all these all these 5 independent variables are you know connected to US economy as a result the energy consumptions and nuclear generations, coal production, gas production and fuel rate may have some kind of you know internal link.

So, what will have what will have here? So, you will we will find out how long there you know related to each others. So, technically in order to establish this let us see the case. So, in the first instance if you actually you know build the model. So, we will have a system like this.

(Refer Slide Time: 05:55)



Predict Crude Oil Production

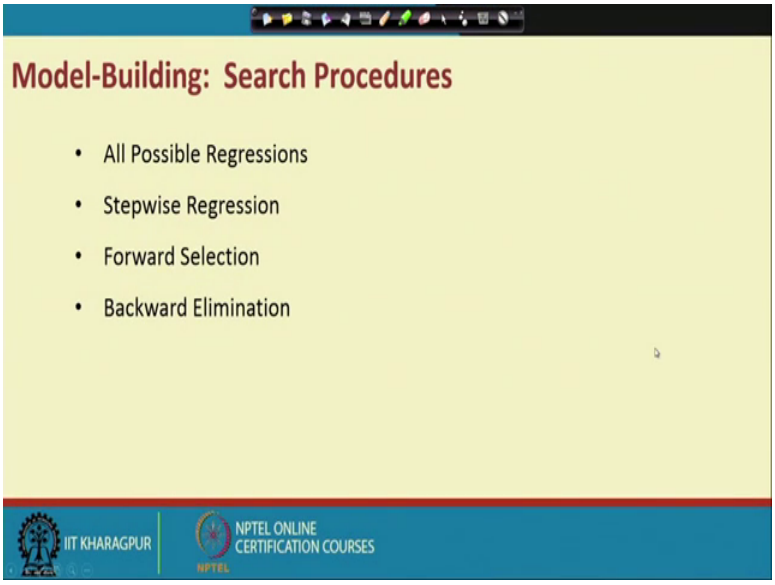
Y	World Crude Oil Production
X_1	U.S. Energy Consumption
X_2	U.S. Nuclear Generation
X_3	U.S. Coal Production
X_4	U.S. Dry Gas Production
X_5	U.S. Fuel Rate for Autos

Y	X_1	X_2	X_3	X_4	X_5
55.7	74.3	83.5	598.6	21.7	13.30
55.7	72.5	114.0	610.0	20.7	13.42
52.8	70.5	172.5	654.6	19.2	13.52
57.3	74.4	191.1	684.9	19.1	13.53
59.7	76.3	250.9	697.2	19.2	13.80
60.2	78.1	276.4	670.2	19.1	14.04
62.7	78.9	255.2	781.1	19.7	14.41
59.6	76.0	251.1	829.7	19.4	15.46
56.1	74.0	272.7	823.8	19.2	15.94
53.5	70.8	282.8	838.1	17.8	16.65
53.3	70.5	293.7	782.1	16.1	17.14
54.5	74.1	327.6	895.9	17.5	17.83
54.0	74.0	383.7	883.6	16.5	18.20
56.2	74.3	414.0	890.3	16.1	18.27
56.7	76.9	455.3	918.8	16.6	19.20
58.7	80.2	527.0	950.3	17.1	19.87
59.9	81.3	529.4	980.7	17.3	20.31
60.6	81.3	576.9	1029.1	17.8	21.02
60.2	81.1	612.6	996.0	17.7	21.69
60.2	82.1	618.8	997.5	17.8	21.68
60.6	83.9	610.3	945.4	18.2	21.04
60.9	85.6	640.4	1033.5	18.9	21.48

$Y = f(X_1\ X_2\ X_3\ X_4, X_5)$

So, Y equal to function of X 1 X 2 X 3 X 4 and X 5, but you know there is a high chance that you know all these 5 variables you may not be in the final setup because of this multicollinearity problem. So, if you could actually include all these 5 variables without multicollinearity problem, then this is excellent, but if that is not the case. So, we will find out the scenario or the good models where more number of independent variables will be there, and they are you know free from the multicollinearity problem. So, as a result, so, we will have a different kind of you know you know shapes through which you can actually recap a particular you know items depending upon our you know requirement.

(Refer Slide Time: 06:45)



Model-Building: Search Procedures

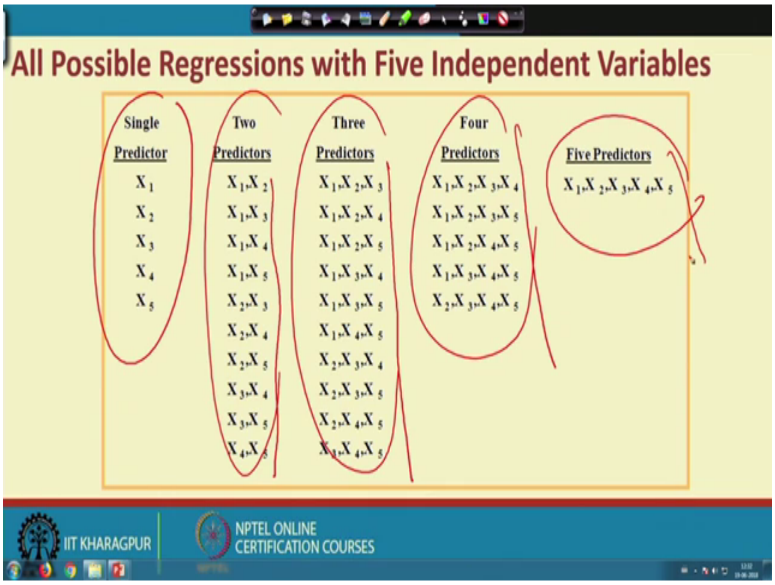
- All Possible Regressions
- Stepwise Regression
- Forward Selection
- Backward Elimination

So, ultimately we have a couple of you know structure. So, far as a multicollinearity concern first all possible regressions, then stepwise regression ; that means, technically out of 5 variables if you say all possible regression means its Y with you know all independent variable simultaneously. Now the moment you will find the all these independent variables are not actually fit for the you know model building, then we will go for the you know you know stepwise case. So, where we will have a where we will have a kind of you know structure that will it take you, take you to the case where it will be multi you know free from you know multicollinearity problem.

So that means, technically. So, we have a structure called as a stepwise regression. So, it will go step by step, to check you know which are the important variables are you know good for these particular you know predictions, and that to avoiding the multicollinearity problem. So, stepwise regression has a 2 options, forward selection and backward elimination. In the forward selection process you start with you know the most important variables, then the second most third most and every times we have to check the model accuracy and the correlation among the regression. And finally, we will freeze a particular you know you know structures where, all these variables will be there in the system and that too have no multicollinearity problem.

Similarly, in the backward elimination process, you start with the full model then the most you know least variable which is not affected drastically. So, that will be dropped, then subsequently to drop 1 by one until we get a model which is actually a good for the estimation prediction and forecasting, and without any multicollinearity problem. This is actually the last part of the game where we need actually final models, which is free from multicollinearity problem. But before that what we will do? We will we like to check you know what are the ways you can actually detect the multicollinearity, how is the level of multicollinearity, what is the consequence of multicollinearity then how to look for the solutions course this is one of the solution, but this is not the ultimate solution.

(Refer Slide Time: 09:17)



All Possible Regressions with Five Independent Variables

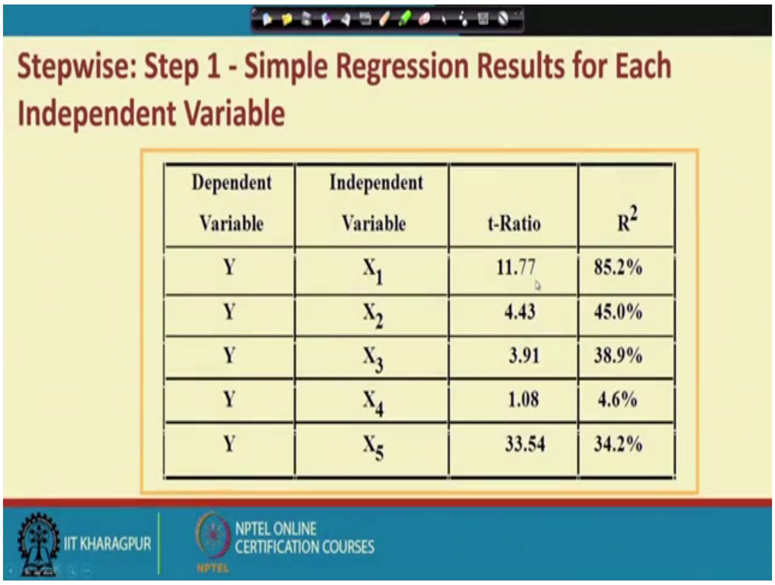
Single Predictor	Two Predictors	Three Predictors	Four Predictors	Five Predictors
X_1	X_1,X_2	X_1,X_2,X_3	X_1,X_2,X_3,X_4	X_1,X_2,X_3,X_4,X_5
X_2	X_1,X_3	X_1,X_2,X_4	X_1,X_2,X_3,X_5	
X_3	X_1,X_4	X_1,X_2,X_5	X_1,X_2,X_4,X_5	
X_4	X_1,X_5	X_1,X_3,X_4	X_1,X_3,X_4,X_5	
X_5	X_2,X_3	X_1,X_3,X_5	X_2,X_3,X_4,X_5	
	X_2,X_4	X_1,X_4,X_5		
	X_2,X_5	X_2,X_3,X_4		
	X_3,X_4	X_2,X_3,X_5		
	X_3,X_5	X_2,X_4,X_5		
	X_4,X_5	X_3,X_4,X_5		

So, let us see what are the ways we can actually discuss. Now, corresponding to these if you know you know multiple regression problem, where we have dependent variable with the 5 independent variable. So, we have a following kind of you know modelling structures; Y with single predictor Y with 2 predictors Y with 3 predictors, Y with four predictors and Y with the 5 predictors this is when we will go for you know model design you can have like this.

But ultimately if you start with you know first to last so; that means, this is the single predictor with you know Y upon X 1, Y upon X 2, Y upon; that means, there are 5 such instances similarly here if it is 2 predictors, then we have a couple of instances 3 predictors we have also couple of instances, four predictors we have couple of instances and 5 predictors, we have couple of instances; that means, all together we have all such you know possibilities. Then out of which only one possibility final we will pick up to predict the oil productions subject to all these independent variables relating to us market.

So now, how we can do that? So, let us see so; that means, technically before you go to this you know process.

(Refer Slide Time: 10:35)



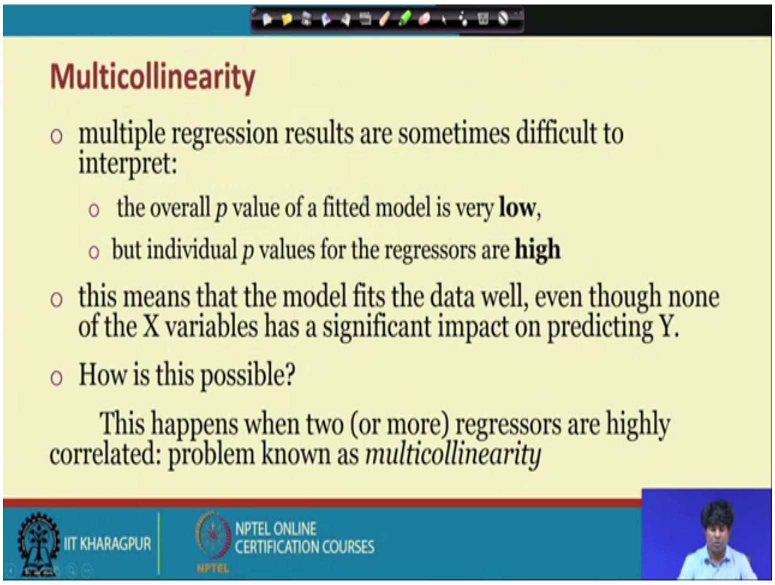
Stepwise: Step 1 - Simple Regression Results for Each Independent Variable

Dependent Variable	Independent Variable	t-Ratio	R^2
Y	X_1	11.77	85.2%
Y	X_2	4.43	45.0%
Y	X_3	3.91	38.9%
Y	X_4	1.08	4.6%
Y	X_5	33.54	34.2%

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, the stepwise process is like that you know regress Y with the all these independent variables 1 by one, and you will find you know most important variable is the X 1 that is a energy consumption and then followed by a the X 2, then X 3, and then X 5, and then X 4. So, like you know they have a different impact, they are not actually equal impact. So, this is as usual you know structure of the regression you cannot find uniform impact of you know all these independent variable some will import you know very high some impact very low so, likewise this is the reality.

(Refer Slide Time: 11:15)



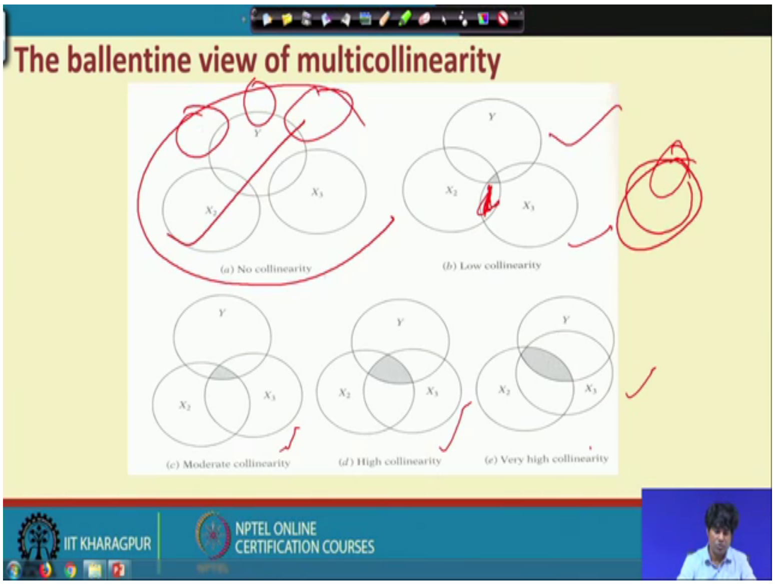
Multicollinearity

- multiple regression results are sometimes difficult to interpret:
  - the overall *p* value of a fitted model is very **low**,
  - but individual *p* values for the regressors are **high**
- this means that the model fits the data well, even though none of the X variables has a significant impact on predicting Y.
- How is this possible?

This happens when two (or more) regressors are highly correlated: problem known as *multicollinearity*

So, after knowing all these details, let us come to the multicollinearity problem again. So, it is the question of you know linear relationship among the regressors. So, and a game of independent variables, the relationship is linear in natures, multivariate framework, then we are looking for the degree of you know association. Then we will we will try to check whether the degree of relationship or degree of association between these independent variables are statistically significant or not statistically significant, that is how the multicollinearity come into the picture to select a model which can be free from the multicollinearity issue.

(Refer Slide Time: 11:55)



The ballentine view of multicollinearity

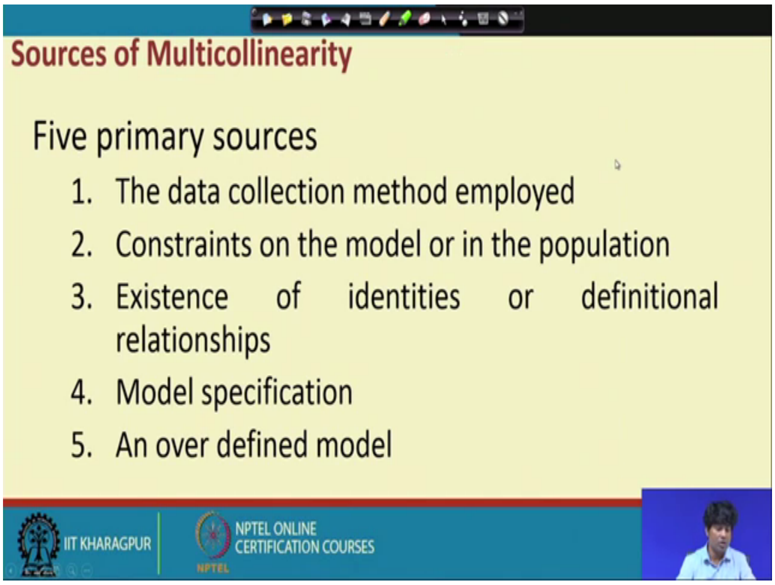
Y
X_2
X_3
(a) No collinearity
(b) Low collinearity
(c) Moderate collinearity
(d) High collinearity
(e) Very high collinearity
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, these are the possible you know steps you can find in the real life scenario, and if you go through this you know steps this is the case where we have no multicollinearity and this is the best, and this is low multicollinearity, and this is a little bit moderate, this is a high and this is actually very high multicollinearity. So, when we called as you know all the correlation coefficients among the independent variables are 0 then; that means, technically you can have like this. If you say add another independent variable, then it can be like this then it can be like this, it can be like this so; that means, there completely independent.

But now in this case you will find it is between Y and X 2 X 3 somehow there is slight relationship between X 2 and X 3. So, this is actually this may be go ead with the prediction because it has low multicollinearity problem, but this case it is actually having high multicollinearity issue you can create also very big one. So, like this is actually severe problems of muticollinearity so; that means, technically after plotting or you know looking like this. So, you can guess you know what is the level of multicollinearity of course, we can quantify, the level of multicollinearity problem by noting the correlation coefficient, close to 1 is a high multicollinearity, close to 0 is very low multicollinearity.

So, when it is close to 0 then this is good for the model building and if it is equal to 1 it is actually not good for the model building.

(Refer Slide Time: 13:31)



Sources of Multicollinearity

Five primary sources

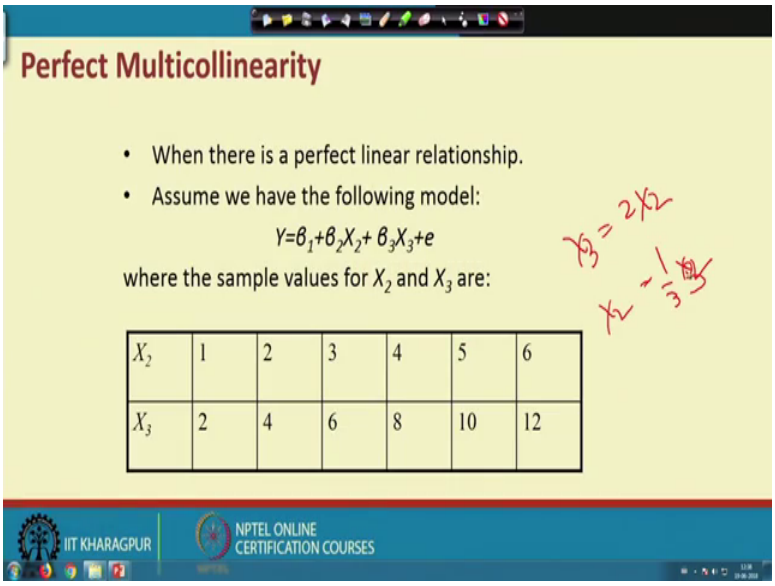
1. The data collection method employed
2. Constraints on the model or in the population
3. Existence of identities or definitional relationships
4. Model specification
5. An over defined model

So, likewise we will have the understanding and then what are the actually reasons through which you can actually expect that there will be a multicollinearity? Sometimes, you know as usual the variables are you know by default there related to each other, sometimes the model specification maybe wrongs. So, the sample selection may be wrong, then you know sometimes there is a question of over identify model, and sometimes low sample size or excessive sample size may also create such kind of you know problem.

So, we have to actually find out a optimal kind of you know environment where, we have the estimated model and which is actually free from all this kind of you know issues. You know in the first instance it is you have to find out the existence of you know multicollinearity, then you think about the reasons at the kind of you know relations. The reasons a reasons of you know having multicollinearity is not so, important ultimately it is in the systems. So, you have to you know clean the particular you know system, otherwise you know you cannot go ead with the prediction and forecasting.

So, what are the sources it is coming that is not more important, what is important you know check the level of multicollinearity, and think about the solution procedures, that is more important. So, per as you know solving any engineering problems with respect to the prediction and forecasting.

(Refer Slide Time: 14:57)



Perfect Multicollinearity

- When there is a perfect linear relationship.
- Assume we have the following model:

$$Y=\beta_1+\beta_2X_2+\beta_3X_3+e$$

where the sample values for X_2 and X_3 are:

X_2	1	2	3	4	5	6
X_3	2	4	6	8	10	12

$X_3 = 2X_2$

$X_2 = \frac{1}{3}X_3$

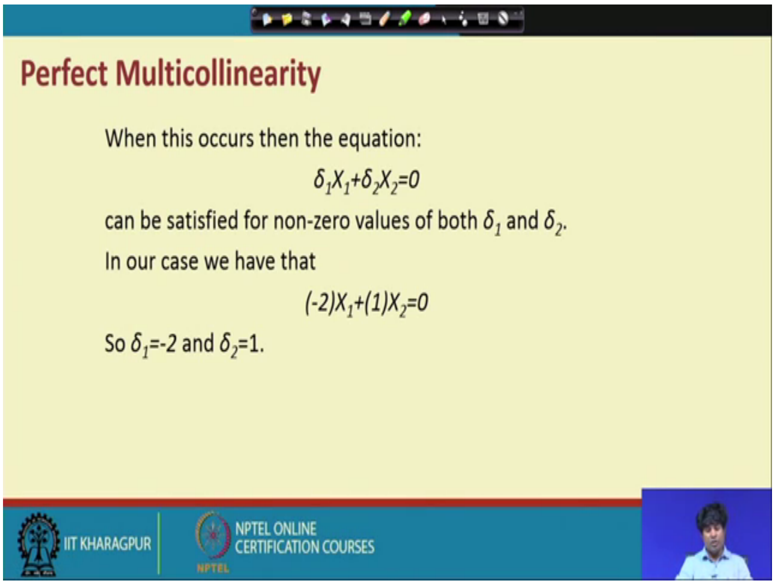
So, this is one way of you know saying perfect multicollinearity. So, here you see X 2 and X 3. So, technically X 2 is nothing, but you know 2 of X 3 is you know nothing, but 2 of X 2 so; that means, in this case. So, X 3 equal to 2 of X 2 more or X 2 equal to 1 by 3 X 2 so; that means, ; that means, technically same data points are you know are repeated with you know different levels by multiplying 2.

So; that means, a technically X 2 and X 3 are same. So, in the if you know model with you know Y X 2 and X 3 simultaneously, then you do not find any result at all. So, sometimes you know if you use any software, software will be give you some kind of you know error you know signal. So, since its actually very easily you can detect or you can recognize. So, in the first instance either it keep X 2 or X 3, but you cannot keep actually X 2 X 3 simultaneously because this is the case of you know perfect multicollinearity. So, you have to remove any one of these 2 variables.

So, now mathematically without knowing X 2 and X 3 you can draw either X 2 or x 3, but theoretically or you know any kind of you know real life problem engineering problem. So, before dropping X 2 or X 3 you have to think which one is more important, which is which one is more relevant to this particular problem then you can keep that one send the other one you can actually so; that means, prime importance you can . So, the level of you know inclusions not only through statistical criteria, but also theoretical you have to check the prime reasons or you know prime impact you know both theoretically and statistical and then finally, you have to freeze as per the particular you know requirement.

Ultimately the structure will be like this and so, this is actually the case perfect multicollinearity case.

(Refer Slide Time: 17:00)



Perfect Multicollinearity

When this occurs then the equation:

$$\delta_1 X_1 + \delta_2 X_2 = 0$$

can be satisfied for non-zero values of both δ_1 and δ_2.

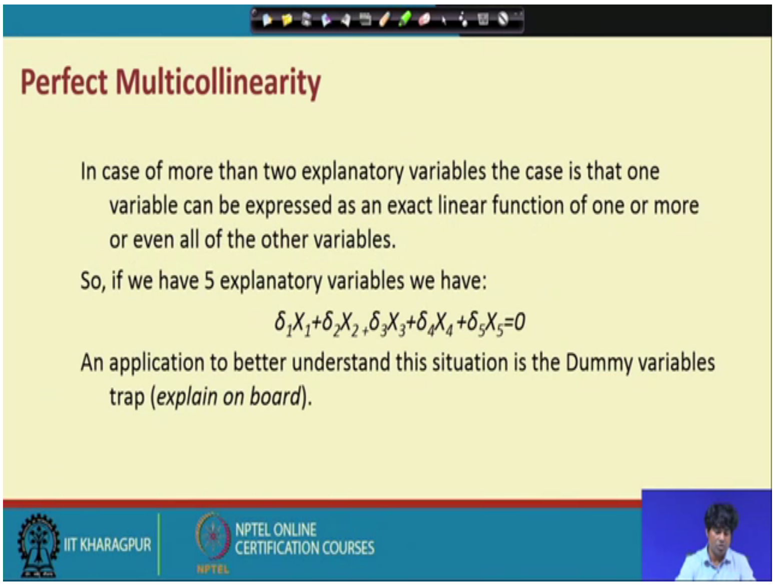
In our case we have that

$$(-2)X_1 + (1)X_2 = 0$$

So $\delta_1 = -2$ and $\delta_2 = 1$.

And in that case if we you establish linear relationship, then the then this will be we declared that you know both the variables are linearly dependent for. So, as a result you cannot actually use this model for the case of you know model building or model estimations.

(Refer Slide Time: 17:13)



Perfect Multicollinearity

In case of more than two explanatory variables the case is that one variable can be expressed as an exact linear function of one or more or even all of the other variables.

So, if we have 5 explanatory variables we have:

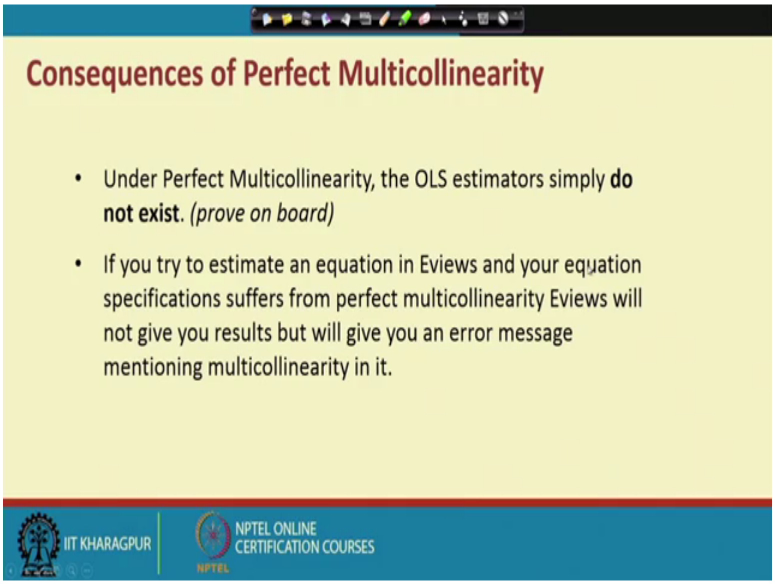
$$\delta_1 X_1 + \delta_2 X_{2\,+} \delta_3 X_3 + \delta_4 X_4 + \delta_5 X_5 = 0$$

An application to better understand this situation is the Dummy variables trap (*explain on board*).

Similarly if 5 explanatory variables, the structure will be like this then you check whether these coefficients are you know statistically significant or not.

For instance if you have actually 5 variables, you can connect actually X 2 upon X 3 X 2 upon X 4, and then check these coefficients are statistically significant. That means, knowing the existence of you know multicollinearity either you can apply covariance or correlation or even you know regressions, these are the standard you know structure through which you can detect of course, at the upper levels some checks are there, which can give you little bit hint to know the existence of multiocollinearity problem in the estimation process and this is the case of you know consequences.

(Refer Slide Time: 18:03)



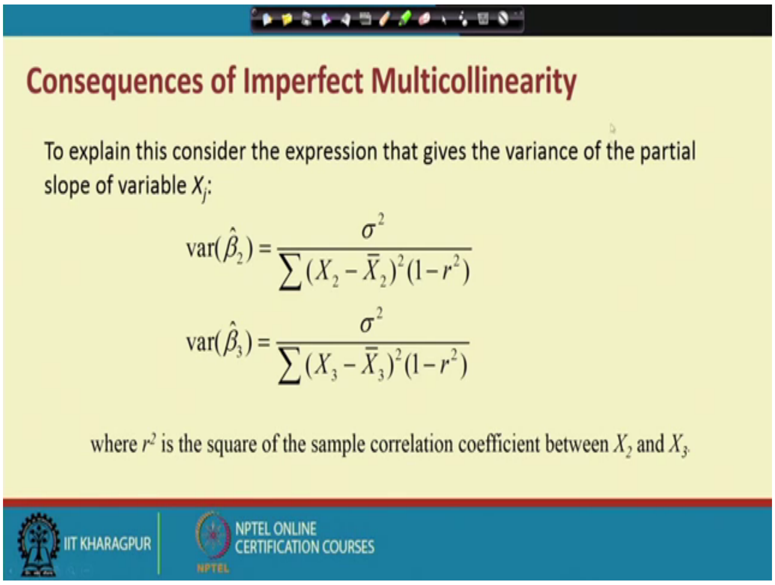
Consequences of Perfect Multicollinearity

- Under Perfect Multicollinearity, the OLS estimators simply **do not exist**. *(prove on board)*
- If you try to estimate an equation in Eviews and your equation specifications suffers from perfect multicollinearity Eviews will not give you results but will give you an error message mentioning multicollinearity in it.

When there is a multicollinearity problem. So, the here o l s estimators are not actually blue, that is what we have already discussed best linear unbiased estimator. So, usually in the presence of multicollinearity, your standard error will be very high and make the coefficient indeter[minate]- you know you know you know indeterminate or something like you know inconsistent, while you know building models or use the model for the prediction and forecasting. So, well, so, ultimately multicollinearity is a serious problem, in the case of you know regression modelling.

So, these are all the case of you know difference.

(Refer Slide Time: 18:45)



Consequences of Imperfect Multicollinearity

To explain this consider the expression that gives the variance of the partial slope of variable X_j:

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum (X_2 - \bar{X}_2)^2 (1 - r^2)}$$

$$\operatorname{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum (X_3 - \bar{X}_3)^2 (1 - r^2)}$$

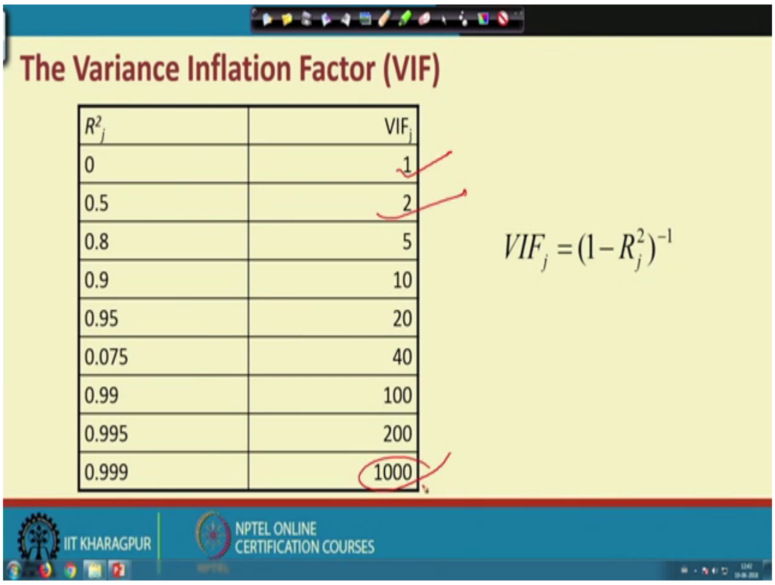
where r^2 is the square of the sample correlation coefficient between X_2 and X_3.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And usually when we try to test the particular variable, the variability part or the standard error of these particular you know estimators will be very high, as a result the t statistic of this coefficient will be indeterminate.

As a result you know it is clear, that you know multicollinearity having multicollinearity, you can get the model you know estimator model reliable one. So, you have to be very careful how you have to deal with the situation.

(Refer Slide Time: 19:15)



The Variance Inflation Factor (VIF)

R^2_j	VIF_j
0	1
0.5	2
0.8	5
0.9	10
0.95	20
0.075	40
0.99	100
0.995	200
0.999	1000

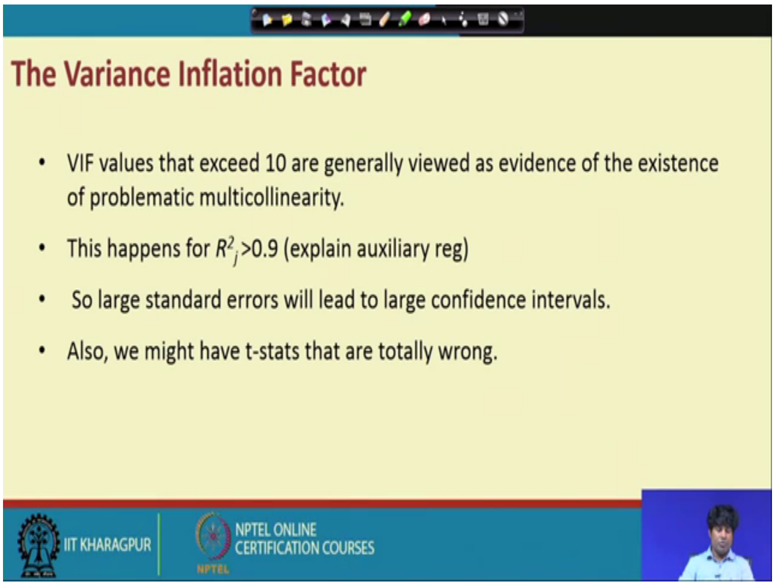
$$VIF_j = (1 - R_j^2)^{-1}$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And there are lots of other indicator through which you can check the multicollinearity, one of one of such indicator is VIF Variance Inflation Factors. It starts with you know 0 to 1. So, it depends upon you know 1 by 1 minus R square that is the coefficient of determination and in this case you will see here, you will see here the kind of you know structures the range it will start with the simply a R square equal to 0 VIF equal to 1 and R square equal to 0.5 VIF equal to 2 you know 2 so; that means, high variation inflation factor. So, it will give you know high multicollinearity problems.

So, this is actually it is a kind of you know perfect and severe multicollinearity problem. So, if you close towards you know R square 0. So, then there will be less you know multicollinearity problem. So, this is how actually you have to check; that means, technically the multicollinearity problem can be check through VIF, with outer knowing the intercept etcetera. So, this will give little bit signal, but ultimately which particular variable will be having or which particular pair will be statistically significant and producing high multicollinearity for that you have to go you know either through correlation matrix or you know simple linear regression or you know covariance approach.

(Refer Slide Time: 20:41)



The Variance Inflation Factor

- VIF values that exceed 10 are generally viewed as evidence of the existence of problematic multicollinearity.
- This happens for R^2_j>0.9 (explain auxiliary reg)
- So large standard errors will lead to large confidence intervals.
- Also, we might have t-stats that are totally wrong.

So, ultimately this clays that you know you know this happens for R square 0.9; so that means, technically VIF values that exceed 10 are generally viewed as evidence of you know existence some problematic multicollinearity; that means, technical what we have already check here. So, this side onward, so, you have actually severe multicollinearity, then degree of multicollinearity low low low low and this is very low and when VIF equal to 1, then there is no multicollinearity problem. So, when it crosses 10 it will have a severe multicollinearity problem.

(Refer Slide Time: 21:15)

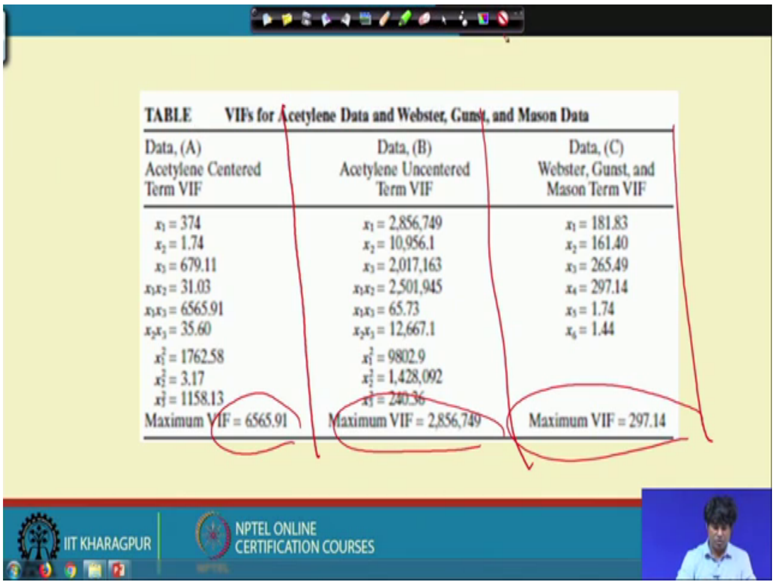


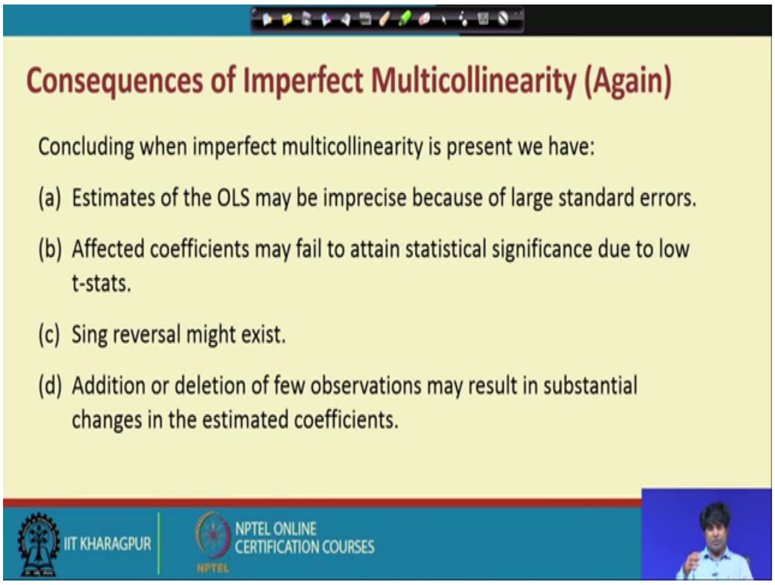
TABLE VIFs for Acetylene Data and Webster, Gunst, and Mason Data

Data, (A) Acetylene Centered Term VIF	Data, (B) Acetylene Uncentered Term VIF	Data, (C) Webster, Gunst, and Mason Term VIF
$x_1 = 374$	$x_1 = 2{,}856{,}749$	$x_1 = 181.83$
$x_2 = 1.74$	$x_2 = 10{,}956.1$	$x_2 = 161.40$
$x_3 = 679.11$	$x_3 = 2{,}017{,}163$	$x_3 = 265.49$
$x_1x_2 = 31.03$	$x_1x_2 = 2{,}501{,}945$	$x_4 = 297.14$
$x_1x_3 = 6565.91$	$x_1x_3 = 65.73$	$x_5 = 1.74$
$x_2x_3 = 35.60$	$x_2x_3 = 12{,}667.1$	$x_6 = 1.44$
$x_1^2 = 1762.58$	$x_1^2 = 9802.9$	
$x_2^2 = 3.17$	$x_2^2 = 1{,}428{,}092$	
$x_3^2 = 1158.13$	$x_3^2 = 240.36$	
Maximum VIF = 6565.91	Maximum VIF = 2,856,749	Maximum VIF = 297.14

And this is the real life examples, and we have a different kind of you know setups in this case there are 3 such instances of course, we will solve a particular problem through software, but in the mean times if you look here, then in the first model VIF is coming 0 you know 6565 and here we have VIF is 2856 here it is actually 297.

So; that means, different level of VI factors when in all the cases the model is actually having multicollinearity problem. This is the clear cut signal and of course, you can actually check through various ways, to justify the you know authenticity of this particular you know problem and, but ultimately if the smoke is coming in one particular mechanism; obviously, you know it will also appear in other such you know instances. So, likewise. So, you can be very careful, how you have to deal such scenario and ok.

(Refer Slide Time: 22:22)



Consequences of Imperfect Multicollinearity (Again)

Concluding when imperfect multicollinearity is present we have:

(a) Estimates of the OLS may be imprecise because of large standard errors.

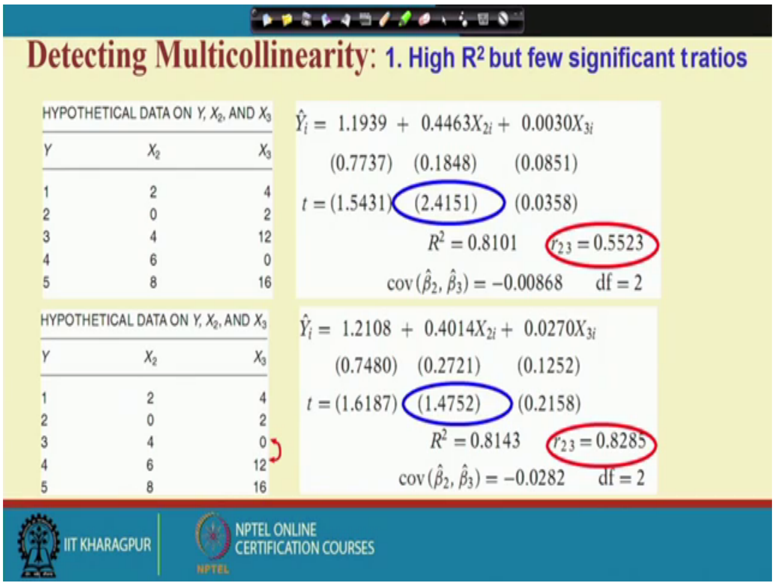
(b) Affected coefficients may fail to attain statistical significance due to low t-stats.

(c) Sing reversal might exist.

(d) Addition or deletion of few observations may result in substantial changes in the estimated coefficients.

So, far as a consequence of you know imperfect multicollinearity is concerned. So, estimators you know will be efficient consistent. So, what will it try to do? So, we like to actually bring a situation where the standard error or variance of these parameters will be very low, as a result the statistical significance of the parameter will be accurate or you know consistent and then it will be statistically significant.

(Refer Slide Time: 22:53)



Detecting Multicollinearity: 1. High R² but few significant tratios

HYPOTHETICAL DATA ON Y, X_2, AND X_3

Y	X_2	X_3
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

$$\hat{Y}_i = 1.1939 + 0.4463X_{2i} + 0.0030X_{3i}$$
$$(0.7737) \quad (0.1848) \quad (0.0851)$$
$$t = (1.5431) \quad (2.4151) \quad (0.0358)$$
$$R^2 = 0.8101 \qquad r_{23} = 0.5523$$
$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.00868 \qquad \text{df} = 2$$

HYPOTHETICAL DATA ON Y, X_2, AND X_3

Y	X_2	X_3
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

$$\hat{Y}_i = 1.2108 + 0.4014X_{2i} + 0.0270X_{3i}$$
$$(0.7480) \quad (0.2721) \quad (0.1252)$$
$$t = (1.6187) \quad (1.4752) \quad (0.2158)$$
$$R^2 = 0.8143 \qquad r_{23} = 0.8285$$
$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.0282 \qquad \text{df} = 2$$

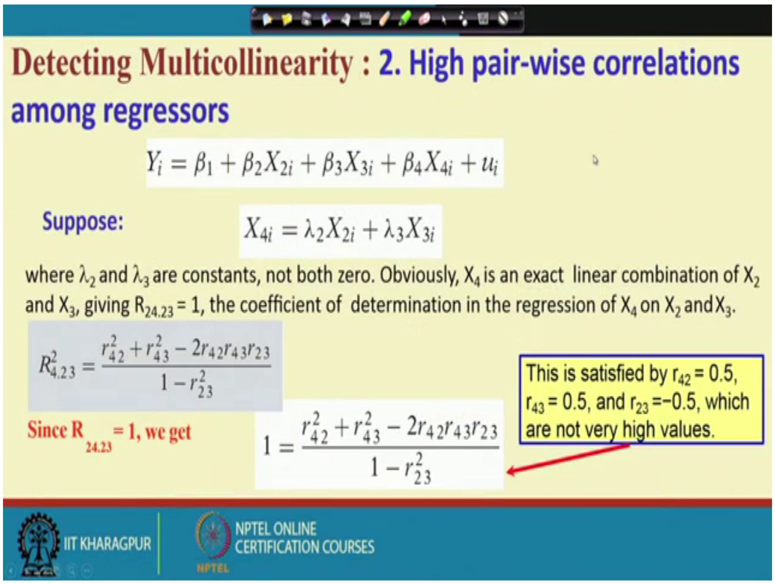
So, ultimately we have to bring such kind of you know environment. Even the even you like to drop some variables does not matter, but ultimately your model should be free from all kind of you know obstacles.

So, far as a detection criteria is concerned, so, we already gone we have already highlighted the covariance mechanism; correlation mechanism regression mechanism, then the VIF mechanism and some of the other criteria like here very simple way to you know check the model and give the answers; that means, sometimes what happens? There is a indicator called as a coefficient determinant R square, and the beta coefficient that is the coefficient against all these independent variable. So, the model requirement is like that or you can say that the model is absolutely fine for the prediction and forecasting, when R square is high and in the same time, all the parameters are statistically significant; that means, all the beta coefficient will be statistically significant.

So, now if you few R square is high few parameters are statistically significant and others are not significant, then this may be the due to multicollinearity problem reverse is also true. So, when you know most of the variables or parameters are significant and R square is low, this is also actually bed signal. So that means, the signal will be very good and green. So, when R square is a high most of the variables are statistically significant or vice versa is also true when most of the variables are statistically significant R square should be very high. So, if that is not the case then this will be create a multicollinearity problem.

Or else if low R square then the significance of the parameters will be at the low level, for instance out of 5 only 2 or 3 can be statistically significant and that too maybe at the 10 percent level. But when most of the variables are statistically significant and R square is not very high, then this leads to a you know problem of you know multicollnearity problem.

(Refer Slide Time: 24:57)



Detecting Multicollinearity : 2. High pair-wise correlations among regressors

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

Suppose: $X_{4i} = \lambda_2 X_{2i} + \lambda_3 X_{3i}$

where λ_2 and λ_3 are constants, not both zero. Obviously, X_4 is an exact linear combination of X_2 and X_3, giving $R_{24.23} = 1$, the coefficient of determination in the regression of X_4 on X_2 and X_3.

$$R^2_{4.23} = \frac{r^2_{42} + r^2_{43} - 2r_{42}r_{43}r_{23}}{1 - r^2_{23}}$$

Since $R_{24.23} = 1$, we get

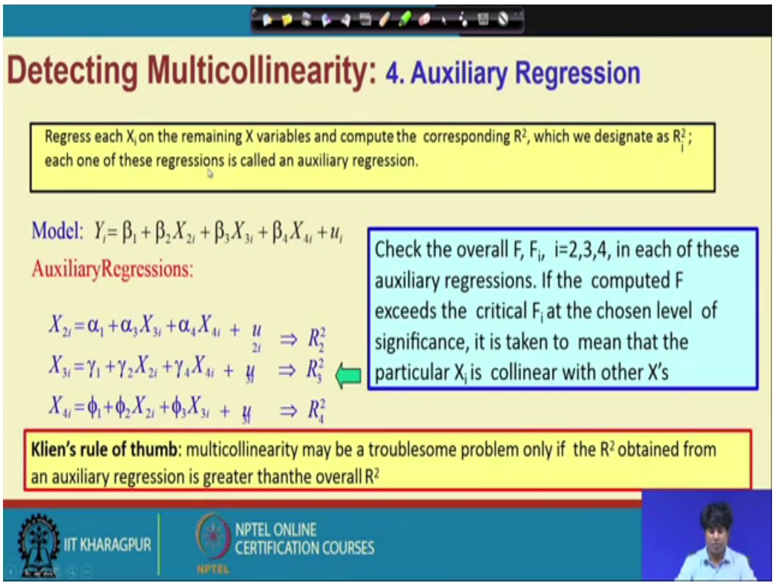
$$1 = \frac{r^2_{42} + r^2_{43} - 2r_{42}r_{43}r_{23}}{1 - r^2_{23}}$$

This is satisfied by $r_{42} = 0.5$, $r_{43} = 0.5$, and $r_{23} = -0.5$, which are not very high values.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Likewise you know ok. So, what I have already mentioned. So, you can regress one independent variable with the rest of the independent variables, you can use simple correlation, you can use partial correlation. So, then you can find the level of that means, high pair wise correlation among the regressor will give you the multicollinearity issue then ok.

(Refer Slide Time: 25:24)



Detecting Multicollinearity: 4. Auxiliary Regression

Regress each X_i on the remaining X variables and compute the corresponding R^2, which we designate as R^2_i; each one of these regressions is called an auxiliary regression.

Model: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

AuxiliaryRegressions:

$$X_{2i} = \alpha_1 + \alpha_3 X_{3i} + \alpha_4 X_{4i} + u_{2i} \Rightarrow R^2_2$$

$$X_{3i} = \gamma_1 + \gamma_2 X_{2i} + \gamma_4 X_{4i} + u_{3i} \Rightarrow R^2_3$$

$$X_{4i} = \phi_1 + \phi_2 X_{2i} + \phi_3 X_{3i} + u_{3i} \Rightarrow R^2_4$$

Check the overall F, F_i, i=2,3,4, in each of these auxiliary regressions. If the computed F exceeds the critical F_i at the chosen level of significance, it is taken to mean that the particular X_i is collinear with other X's

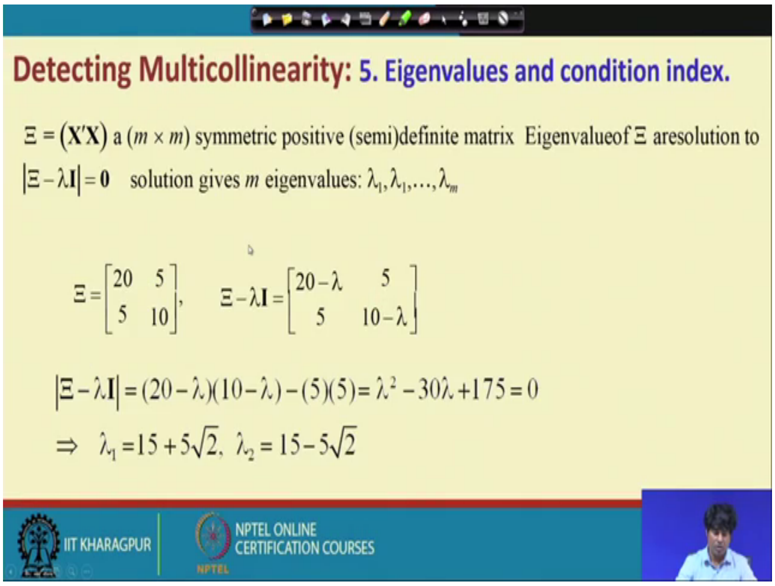
Klien's rule of thumb: multicollinearity may be a troublesome problem only if the R^2 obtained from an auxiliary regression is greater thanthe overall R^2

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is what we know partial correlation and this is what the auxiliary regressions.

So; that means, from these auxiliary regressions you can find out R square and then you can connect with the VIF also. So that means, ultimately we have already highlighted couple of you know different ways, to check the multicollinearity and to validate the estimated models and check whether there in favor of the OLS mechanism or there favor in you know as per the particular you know model requirement.

(Refer Slide Time: 25:44)



Detecting Multicollinearity: 5. Eigenvalues and condition index.

$\Xi = (\mathbf{X'X})$ a $(m \times m)$ symmetric positive (semi)definite matrix Eigenvalueof Ξ aresolution to $|\Xi - \lambda\mathbf{I}| = \mathbf{0}$ solution gives m eigenvalues: $\lambda_1, \lambda_1, \ldots, \lambda_m$

$$\Xi = \begin{bmatrix} 20 & 5 \\ 5 & 10 \end{bmatrix}, \quad \Xi - \lambda\mathbf{I} = \begin{bmatrix} 20-\lambda & 5 \\ 5 & 10-\lambda \end{bmatrix}$$

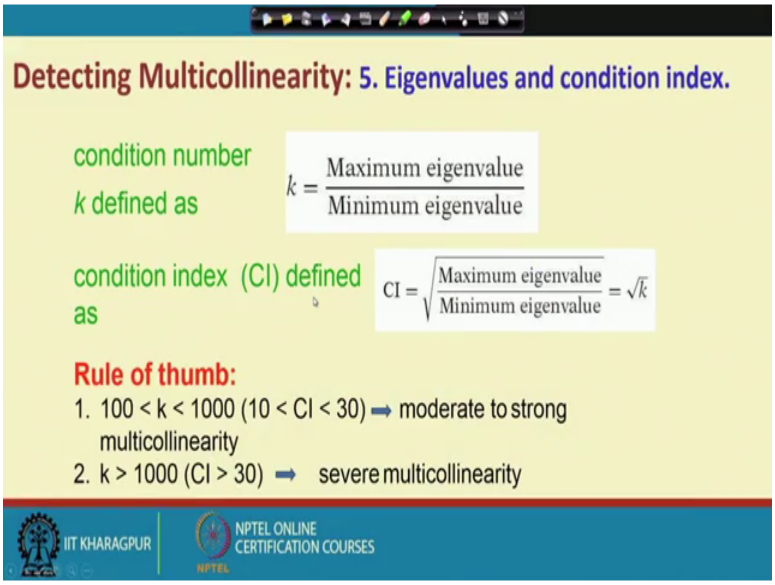
$$|\Xi - \lambda\mathbf{I}| = (20-\lambda)(10-\lambda) - (5)(5) = \lambda^2 - 30\lambda + 175 = 0$$

$$\Rightarrow \quad \lambda_1 = 15 + 5\sqrt{2}, \; \lambda_2 = 15 - 5\sqrt{2}$$

So, through eigenvalue and eigenvector also you can check the multicollinearity problems. So, since we have already mentioned that you know all these independent variables are linearly independent. So that means, if you calculate the correlation matrix if there will linearly independent, then there correlation matrix or value of that particular matrix will be ones; that means, that will take you to the unit matrix.

But if there is multicollinearity problems, then these mo[dels]- you know particular matrix cannot be transfer into the unit matrix. So, this gives the signal that you know there is a somehow multicollinearity problems; that means, technically there is a linear relationship among the regressors.

(Refer Slide Time: 26:35)



Detecting Multicollinearity: 5. Eigenvalues and condition index.

condition number k defined as
$$k = \frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}$$

condition index (CI) defined as
$$\text{CI} = \sqrt{\frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}}} = \sqrt{k}$$

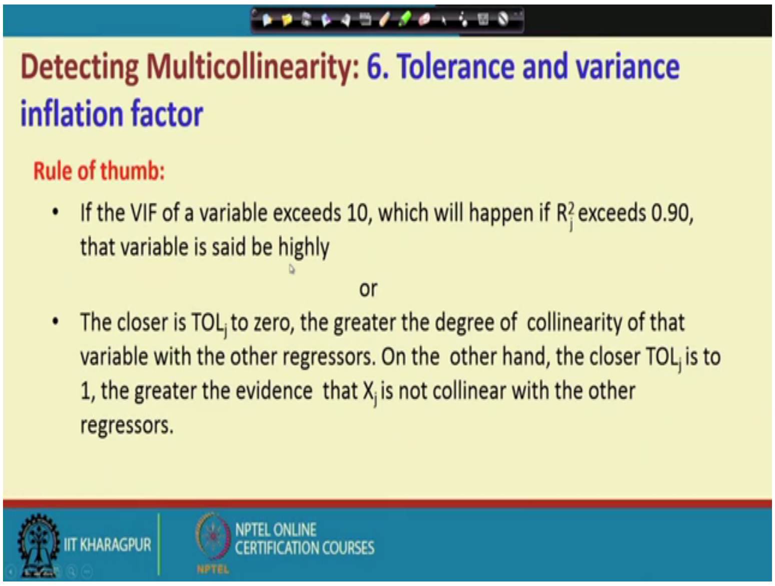
Rule of thumb:

1. $100 < k < 1000$ $(10 < \text{CI} < 30)$ ➡ moderate to strong multicollinearity
2. $k > 1000$ $(\text{CI} > 30)$ ➡ severe multicollinearity

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And again there is a component called as a condition index, which is actually square root of you know maximum eigenvalue by minimum eigenvalue and when the k you know moves in between actually 100 to 1000, then it is moderately you know exist, but when it crosses 1000 then it is a severe multicollinearity; like you know VIF it once it will cross ten then this will be actually having severe problem.

(Refer Slide Time: 27:04)



Detecting Multicollinearity: 6. Tolerance and variance inflation factor

Rule of thumb:

- If the VIF of a variable exceeds 10, which will happen if R_j^2 exceeds 0.90, that variable is said be highly

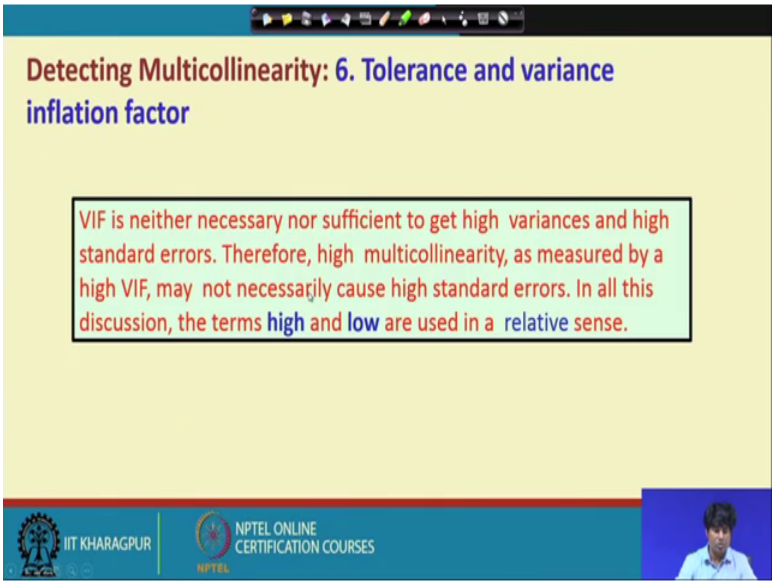
or

- The closer is TOL_j to zero, the greater the degree of collinearity of that variable with the other regressors. On the other hand, the closer TOL_j is to 1, the greater the evidence that X_j is not collinear with the other regressors.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

If it is on the 10, then it will be having less you know less problem for the prediction and forecasting.

(Refer Slide Time: 27:12)

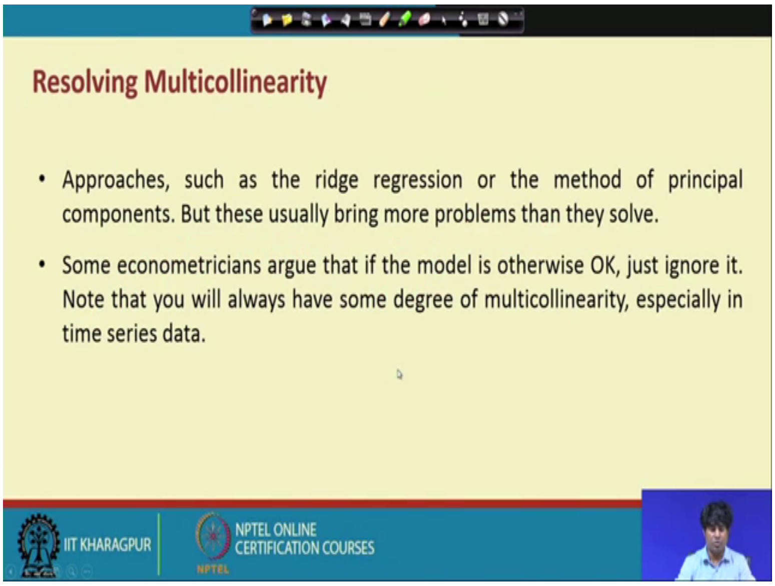


Detecting Multicollinearity: 6. Tolerance and variance inflation factor

VIF is neither necessary nor sufficient to get high variances and high standard errors. Therefore, high multicollinearity, as measured by a high VIF, may not necessarily cause high standard errors. In all this discussion, the terms **high** and **low** are used in a relative sense.

So, this is actually we have already discussed this VIF and that means, technical tolerance vectors. So, the 1 minus R square is the tolerance vector. So, VIF actually calculated by the you know 1 by tolerance and tolerance is the tolerance exactly equal to one minus coefficient of determinate that is R square j.

(Refer Slide Time: 27:30)

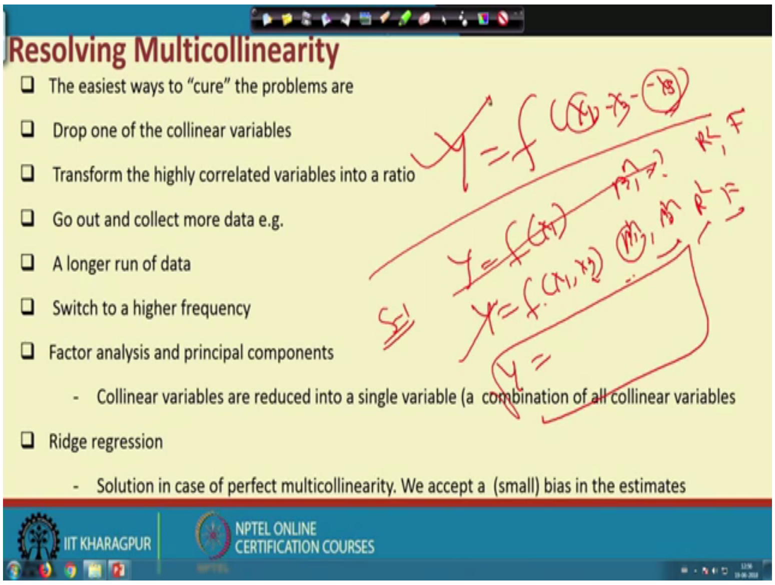


Resolving Multicollinearity

- Approaches, such as the ridge regression or the method of principal components. But these usually bring more problems than they solve.
- Some econometricians argue that if the model is otherwise OK, just ignore it. Note that you will always have some degree of multicollinearity, especially in time series data.

Now, the question is you know how to actually means solve this particular you know multicollinearity. We have already gone through several ways to detect the multicollinearity get to know the consequence and we are looking for the solutions.

(Refer Slide Time: 27:48)



Resolving Multicollinearity

❑ The easiest ways to "cure" the problems are

❑ Drop one of the collinear variables

❑ Transform the highly correlated variables into a ratio

❑ Go out and collect more data e.g.

❑ A longer run of data

❑ Switch to a higher frequency

❑ Factor analysis and principal components

- Collinear variables are reduced into a single variable (a combination of all collinear variables

❑ Ridge regression

- Solution in case of perfect multicollinearity. We accept a (small) bias in the estimates

$Y = f(X_1, X_2, \ldots X_8)$
RK, F
$Y = f(X_1)$
$Y = f(X_1, X_2)$
S:
$Y =$
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

And so, far as a solution is concerned there are multiple solutions we have and one you have to pick up which one is good for your you know model building and model estimations and that too to predict any engineering problem and the kind of you know forecasting.

So, one way to solve the problem is you can simply go by increasing sample size or decrease the sample size, variety of ways actually you can check then you input, you can you know change then you will find output change, then you compare different levels of output and fix a particular model where the a problem is not at all there.

Then simple you know other way is drop the collinear variables and then go ahead with the prediction and forecasting. But dropping a particular variable is not a good strategy, for instance if that particular variable theoretically is very sound and you know having strong impact, then you cannot just drop the variables.

So, you have to find out alternative ways, how to keep that particular variable and that to a model is free from multicollinearity. Then sometimes we will go for you know variable transformations by you know highest difference transformation, low transformations by that way. So, the existence of relationship may be little bit you know compromise or you know reduce.

And you know you try to improve increase the sample size and you can for cross sectional modeling, you can bring different you know sampling so, that you know it will be it will not create any kind of you know such problem. But in the case of time series data so, the existence of relationship will be always there because the data are actually well connected with the time series, that is how the consistency. So, somehow they there will be a relationship, but that relationship should not be statistically significant. Then sometimes you can actually structure the data with you know high frequency . In fact, that is how the kind of you know data restructuring is required either you go for data transformation or somehow you can do the labeling.

So, instead of you know annual data you can go for monthly data or weekly data or quarterly data, somehow this can give you better pictures and the results will be also coming more accurate as per the model requirement or the kind of you know prediction requirement. And sometimes we can use different techniques like you know factor analysis principal component analysis because these are the technique will you usually used when there is a multicollinearity problems. So, usually factor analysis is applied when the structure is k greater than n; that means, number of variables which larger than the sample size.

So, we use factor analysis to transfer such kind of you know environment; that means, instead of using more number of variable, we try to use few number of variables with some kind of you know restriction and the kind of you know mechanism, and this few variables will be by default free from multicollinearity and other errors in the system. But if there is a means this is typically applied, when there is more number of variables in the system.

But if it is a few number of variables are in the systems, then factor analysis may not be used to solve this multicollinearity problem, in that case you look for other different ways; like you know data transformation increase in sample size. Even if you can change the functional form, then you can sometimes drop the collinear variables after knowing the correlation among these regressors.

If there. So, if a particular pair is coming actually very significant for instance let us take closed to 0.99 it is better to drop a particular variables, instead of you know looking for other way of you know solution. So, likewise we you have a different kind of you know atmosphere and then last, but not the least we can use actually the stepwise regression. So, stepwise regression is the last step where you know we will solve the multicollinearity problem, but in the first instance you try to apply all these you know possible techniques by dropping collinear variables, then increase the sample size change functional form, a transfer the data into high frequency data, then in a increase sample size, decrease sample size where different methods like factor analysis apply different structure like you know ridge regressions.

So, these are all various way you can solve the problem then finally, if you could not then you start the stepwise regressions stepwise regressions. So, what I have already mentioned, it has a backward operation and forward operation. In the forward operations you start with the most important variable first, that is the most important independent variable which influence the dependent variable. Then you find out the second most important then you check the reliability of the model every time. Then you have to stop where the reliability will get affected, but instance the starting procedure is a for this problem Y equal to function of like this we what we can do? This is a full model Y equal to function of X to X 5 then you check you know you run the model, and check which particular variable is having more you know high significant a part.

And in the first step of you know step wise. So, Y equal let us say this is nothing, but X 1 only. So, you start with X 1, then see the beta one coefficient value and the statistical significance and the R square value and followed by F statistics, and then you find out another variable which is a high impact after X 1 then let us say it is a X 3. So, then in the next model you start with F X 1 and X 3, then you check beta one coefficients and beta 3 coefficient and R square and F if beta one coefficient if a statistical significant beta 3 significant R square high F significant, then this will be good models you can reject this one.

Again you proceed with another way. So, now, X 1 and X 3 is in. So, X 2 X 4 X 5 is remaining. So, again you have to enter X 2 then check the kind of you know structure if again there improving, then you keep the latest one then remove the previous one. If you know after introducing X 2 if you are no[t]- means X square, if you are not getting good then you can reject this one then what you can do? You can start with you know X 1 X 3 with you know X 4, when you check the environment whether it is good or bad, then agains you compare and you know fix which one is good for you.

So, likewise you can continue till you find a kind of you know structure, which is actually good for the you know forecasting or prediction of the any particular engineering problem, and at the same times it will be free from multicollinearity problem. And in the backward sides you start with you know full models like you know like this then what will you do? You try to drop one after another variable. For instance let us say X 5 is the target variable which is actually having very low impact and that you can get to know with the value coefficient of that X 5 and that though significance of that particular variable.

So, now in the first step of you know backward elimination. So, you drop that variable then you run the model and check the reliability of the model. So obviously, the second step the estimated model will be good one compare to the previous one. Then if it is you know if it is good then you can keep you can reject the previous one, then finally, you come to step 2 where another variable can be dropped. So, you check which one is the least impact again, then you drop that variable ensure that you know remaining of the variables will be statistically significant R square will be high and f will be significant, if that is the case you can go ead with that models and reject the previous one.

So, likewise you will continue then finally, will reach a optimum point where your dependent variable will be predicted with the couple of a independent variables, which are free from multicollinearity problem. That is what the you know requirement of you know OLS mechanism and it is in favor of the OLS assumptions and this is what actually one way called as you know diagnostic check of you know model building.

So, one of the diagnostic check which you can carry whether it is a cross sectional molding or time series modelling is to see whether the relationship among the regressors or independent variable or explanatory variables are you know not related or correlated to each other if that is the case. So, the particular model cannot be you know follow the blue theorem, and cannot be used for you know model estimation or model predictions or you know forecasting.

So, what we will do? So, every time while doing the kind of you know estimations which a be careful to how to check this kind of you know a relationship. So, at the end of the day you must have a estimated model, which is used for prediction and forecasting that is exclusively free from multicollinearity problem. After knowing all these criteria the consequence, the kind of you know detection, the kind of you know solution. I am very sure that you know once you start a particular problem so, you will finally, reach a particular point where you know the estimated model can be you know very efficient to predict and you know forecast, and that too without problems of you know multicollinearity.

But all the multicollinearity is not the diagnostic you know solution or diagnostic check, there are couple of other problems are there like you know heteroscedasticity autocorrelation and some of the other issues. So, which we will discuss in the later stage but in the mean times up to this stage. So, once you go for the multiple regression or multivariate regression. So, we ensure that the final estimated model should be free from multicollinearity problem, that is the existence of linear relationship among the regressors. If that is the case means the model is free from the multicollinearity you can go ead with the prediction to solve the particular engineering problem without any errors and because it is practically you know having the model is a reliable and that too free from multicollinearity without hinting the other components.

But ultimately end of the day, you have to you have to you know check all these you know diagnostic starting with the multicollinearity heteroscedasticity and autocorrelation. But up to this point we are very much sure that you know how to you know fix a particular model, which is free from the multicollinearity problem. In the next class we will discuss the other diagnostics and check the reliability of the model as per the particular you know requirement with this we will stop here.

Thank you, have a nice day.